行動価値関数

⑴ある状態で、特定の行動を選んだ時に、その後どれくらい報酬が期待できるかを示す関数

⑵式：$q_{\pi}(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + rv_{\pi}(s')]$

$q_{\pi}(s,a)$	状態sで行動aを選んだ場合の価値
$\sum_{s'}$	行動aを取った後に遷移する全ての次の状態s'に対して合計
$P(s'\|s,a)$	遷移確率
r	割引率
$v_{\pi}(s')$	次の状態s'の価値

⇓

状態sで行動aを選んだ場合の価値=行動aを取ったあとに遷移するすべての次の状態s'に対しての合計(遷移確率×[即時報酬+割引率×次の状態s'の価値])

行動価値関数とベルマン方程式

⑴ある状態 sで特定の行動 a をとったときに、その後どれくらいの報酬が期待できるかを表す関数

⑵式 : $q_{\pi}(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma \sum_{a'} \pi(a'|s') q_{\pi}(s',a')]$

$q_{\pi}(s,a)$	状態sで行動aを選んだ場合の価値
$\sum_{s'}$	行動aをとったあとに遷移するすべての次の状態s'に対して合計
$P(s'\|s,a)$	遷移確率
$R(s,a,s')$	即時報酬
γ	割引率
$\sum_{a'}$	次の状態s'で選べるすべての行動a'について合計
$\pi(a'\|s')$	次の状態s'において、行動a'を選ぶ確率
$q_{\pi}(s',a')$	次の状態s'で行動a'を選んだ場合の価値

⇓

状態sで行動aを選んだ場合の価値 = 行動aをとったあとに遷移するすべての次の状態s'に対して合計(遷移確率 × [即時報酬 ＋ 割引率 × 次の状態s'で選べるすべての行動a'についての合計 { 次の状態s'において行動a'を選ぶ確率 × 次の状態s'で行動a'を選んだ場合の価値 }])